

**Strategic Plan for the
Unified Production Environment
(DRAFT)**

**NERSC
1994**

Strategic Plan for the Unified Production Environment

Scope of Responsibility

The National Energy Research Supercomputer Center (NERSC) has historically provided a production environment focused on high-capability computing. **The Center has sought to provide (1) the high-end capability platforms and (2) the infrastructure and services necessary to utilize these platforms effectively.** The emphasis to date has rested on a centralized combination of separate services.

Area of Emphasis

The Unified Production Environment (UPE), which we will seek to complete by late 1996, focuses especially on the organization of traditional services into an integrated unit.

Just as a World Wide Web browser presents an intuitive interface through which a user can display information independent of its location, NERSC will present a service interface through which NERSC users will request computing services independent of which machines provide the services. As the researcher employs various services to facilitate progress through all phases of a computational research project, the sense should be that only one environment exists. After logging in, the researcher will see a shared file system regardless of which service is used. At a later stage of the development of the unified environment, the user will have the tools required to effectively utilize the resources through distributed batch and interactive computing. We can view this as a *unification of centralized services*. Ultimately, the Center will seek to integrate the remote user's local environment with the centralized NERSC environment. This could be viewed as a *unification of distributed services*.

An emphasis on service unification is essential at this time because individual components of the service structure (in particular the soon-to-arrive, high-end MPP computational platforms) will require this environment in order to be utilized to fullest potential.

This document provides a strategic foundation for and an overview of its companion document, the Implementation Plan for the Unified Production Environment. Each section of the Implementation Plan addresses one of the elements of the UPE: (1) the Development, Computing, and Assimilation (DCA) environment, (2) the Storage environment, (3) the Local Area Network (LAN), (4) System Administration, and (5) User Services. However, these elements are not and cannot be independent. Their cooperation will define the *Unified Production Environment*. Through the achievement of our four goals as described later in this document, the Center will equalize opportunity for all researchers who compute at the high-end, providing functionally equivalent environments for those with rich and those with only basic local computational support.

Background

The Center, in determining its major goals, must take into account two variables: the first is the technological change occurring in the world, and the second is the users' reactions to this change. The dominant components of **technological change** come from (1) the microprocessor revolution, (2) the adoption of Unix and other standardized protocols (such as X) worldwide, (3) the exponential increase in Wide Area Network (WAN) and LAN bandwidths, and (4) the huge gains in tertiary storage capabilities.

The impact of the microprocessor has already been felt at NERSC. Utilized as a high-end workstation in the form of the Supercomputing Auxiliary Service (SAS), it helped the Center to regain some of the functionality lost through the adoption of UNIX on the supercomputer. SAS provides a rich set of tools and pre-and post processing capabilities. This union between microprocessor and supercomputer represents an initial step to a multicomponent Unified Production Environment (UPE).

Of even greater importance to NERSC is that some of the offspring of the micros have largely surpassed the vector supercomputer in capability. These have evolved into three distinct species: the massively parallel processor (MPP), the symmetric multiprocessor (SMP), and the workstation (or PC) cluster featuring a high performance Asynchronous Transfer Mode (ATM) interconnect. The latter is an interesting, but still unproven technology in the research stage.

The most capable of these three today is the distributed memory, tightly coupled MPP, and it sits at the highest high-end of the capability spectrum. However, the MPP has an immature software base, both from the perspective of the operating system and from the perspective of the tool sets, languages and applications available to the user. To fully utilize the tremendous capability of an MPP will possibly require NERSC to integrate the MPP into two services: (1) complementary capability engines, and (2) the SAS environment.

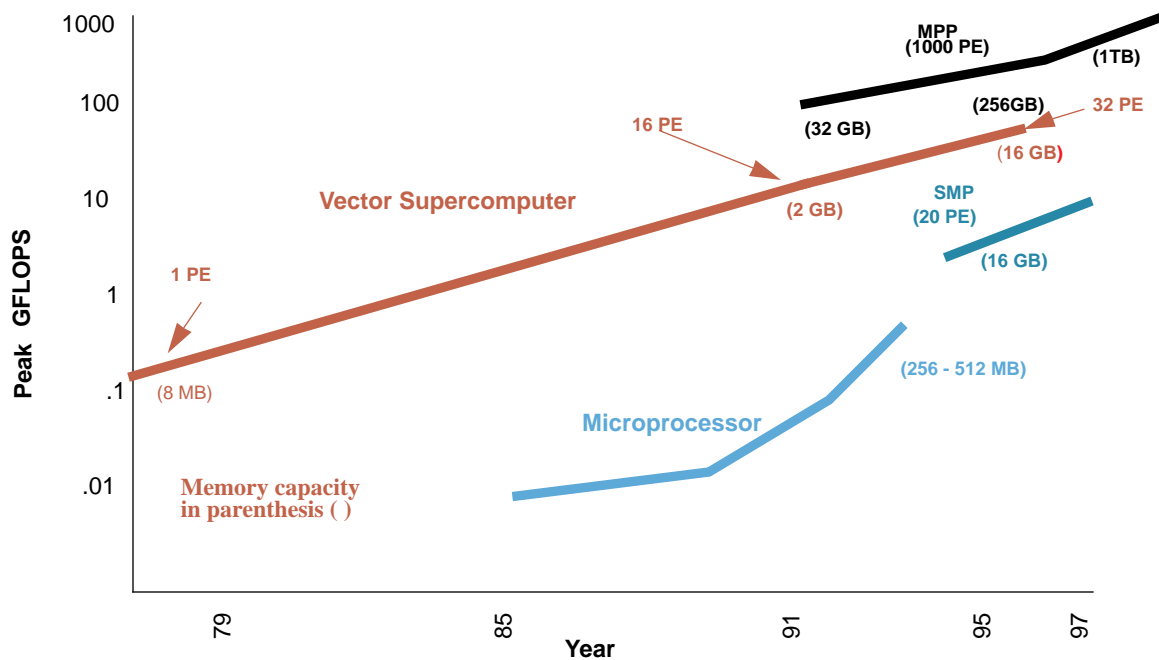
The second offspring is the SMP. From the perspective of the hardware, the SMP differs from the MPP in one essential respect, the memory is not distributed, but shared, just as it is shared on the CRI C90. The SMP will not offer processor counts much beyond 20 to 30 processors, but it will have a rich software environment.

In *Table 1* and *Figure 1* below, we can see how the three systems (MPP, C90 and SMP) might complement one another in a capability context.

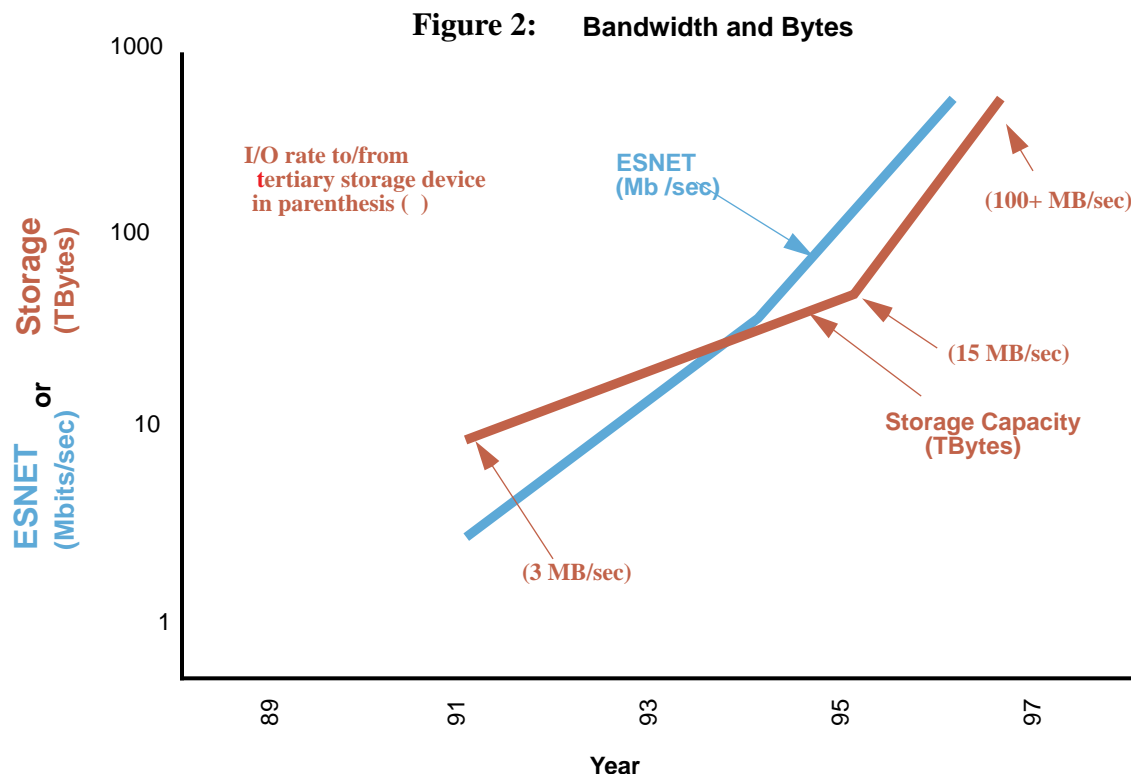
Capability Component	Sustained Speeds (C90/16 units)	Memory (GB)	Development and Assimilation Environment	Apps and Math Libraries	Cost/Perf Ratio (C90/16 units)
C90	1	2 (shared)	good	good	1
MPP '96	4-10	256 (distributed)	minimal	minimal	1/10
SMP '95	.5	16 (shared)	excellent	very good	1/10

Table 1: Attributes of Capability Components

Figure 1: Peak Capability of Various Architectures



In Figure 2 we show the impact of developments in network and storage technologies.



Given this data, one could infer:

- What is considered as computationally intensive supercomputing today (sustained computing at 3-8 gigaflops with a few gigabytes of memory) is nearly achievable on symmetric multiprocessors now. Workstations now have sufficient memory and computational power such that many routine calculations run on the C90 can be successfully completed on these desktop machines. The cost-performance curve is on a very steep improvement slope in time.
- What will be considered as very high-end supercomputing in 1996-7 will be calculations demanding at least 50 gigaflops sustained and/or at least 100 gigabytes of memory¹. The physical models utilized in current codes are designed to put demands on today's machines, not tomorrow's. The next NERSC supercomputer will not be the traditional factor of two more powerful than its predecessor, it will be a factor of ten

1. In the 1996 time frame, NERSC could expect to offer the services of an MPP with at least 512 500 - megaflop processors each possessing at least 256 MB of local memory. A problem utilizing all the processors running at 20% efficiency would then run at 50+ gigaflops sustained. This would represent a reasonable goal for many research codes.

more powerful, and codes taxing this machine will need to be written in different way from in the past. In a sense, it is the quantitative leap that the new machines are making that is part of the problem. Thus, the research community is faced not with an evolutionary change in the way computing is done but in a *revolutionary* change. The transition will be difficult not simply because the programming models are different but because ***the physics incorporated in the codes must be augmented if they are to tax the machine's capability. The MPP allows the physicist to do more complete and more realistic problems.*** The challenge for NERSC is to mitigate this shock by developing an environment which makes at least some of the changes feel evolutionary rather than revolutionary to the scientist.

- The remote user of the high bandwidth wide area network should not sense a difference in accessibility or the feel of locally provided and remotely provided services as the WAN improves. The service tool set could be distributed. Simulation at NERSC coupled with local real-time visualization including feed-back steering should be in the realm of routine computing procedure.
- Storage capacity is evolving commensurately with MPP peak speeds; however, storage I/O lags somewhat and awaits novel software solutions, in particular parallel I/O. High performance storage solutions must be a priority both at NERSC and at similar installations.
- The large bandwidths in the LAN and WAN networks, the huge computational and I/O capability of the high-end MPP, and the near petabyte long term storage capabilities will be largely useless unless the scientist can assimilate the results of his calculations. One must be able both to mine and to visualize the generated data efficiently.

The **users' reactions** to this rapid technological advancement run the gamut from total bewilderment to aggressive leveraging. Some common observations coming from various areas of the research community follow:

- Many current application code developers wish to acquire (or have acquired) private capability such as workstations which provide greater local control over the computational environment, a good code development environment, and a good production environment. They view their research as an intimately interrelated spectrum of efforts, involving use of a variety of codes, some of which tax, and some of which do not tax the most capable NERSC platform.
- Local environments may provide workstations, but users are frustrated when tertiary storage and high performance visualization tools are not available. Proprietary software licenses (such as for engineering codes) are expensive. Local system administration is time consuming. Current software packages and workstation hardware capabilities evolve so quickly that the scientist is forced either to freeze his local environment in order to get work done (and fall behind in vital system administration areas such as security) or to be aggressive and stay current at the expense of research interests.

- In moving from local work environments to the NERSC environment during their computational workday, users require a sense of continuity.
- Because of budgetary constraints, many users with demanding codes are limited to “dumb” X terminals and are essentially reliant on NERSC for *all* computational services.
- Users with the most taxing codes will compute at NERSC *regardless of local resources*, since these resources are not sufficiently capable for grand challenge-scale production runs.

A positive response to these requirements must begin with the premise that the NERSC environment must complement the local environment. This is more easily said than done, because there are as many local environments as there are users. Some of these environments are rich, others are very basic.

Four Basic Goals of the UPE

While the basic mission of NERSC remains essentially unchanged, the new technologies demand that the services offered evolve to become more tightly coupled, sophisticated and global. We have identified four primary goals:

1. **NERSC must continue to serve the needs of high-end computing.** In short, a centrally located MPP within a Unified Production Environment offering adequate support infrastructure (such as disk, tertiary storage and complementary capability platforms) remains a central goal. *Regardless of the improvement in desktop or cluster technology, no local user environment, no matter how rich, can hope to offer peak computational speeds on the order of 300-600 gigaflops or storage capabilities on the order of 200 terabytes by 1996.*
2. **NERSC must continue to offer all the *additional* services required by the high-end user with limited local resources, including the following: (1) code development, computing and assimilation capabilities, (2) archival storage, and (3) information services, consulting and in-depth collaborations.** Information services, consulting and collaborative ventures with NERSC staff are becoming more important because of the complexity in the new programming and assimilation environments. *Without expertise, the capability latent in the hardware will not be realized.*
3. **NERSC must facilitate the users' progression through all the phases of computational research projects by carefully integrating the services described above. That is what is meant by the Unified Production Environment. We must make the revolutionary changes to be faced feel evolutionary.** The MPP is not currently an optimal platform for code development, debugging and short development runs. Ancillary services must be easily accessible, and must be offered on the

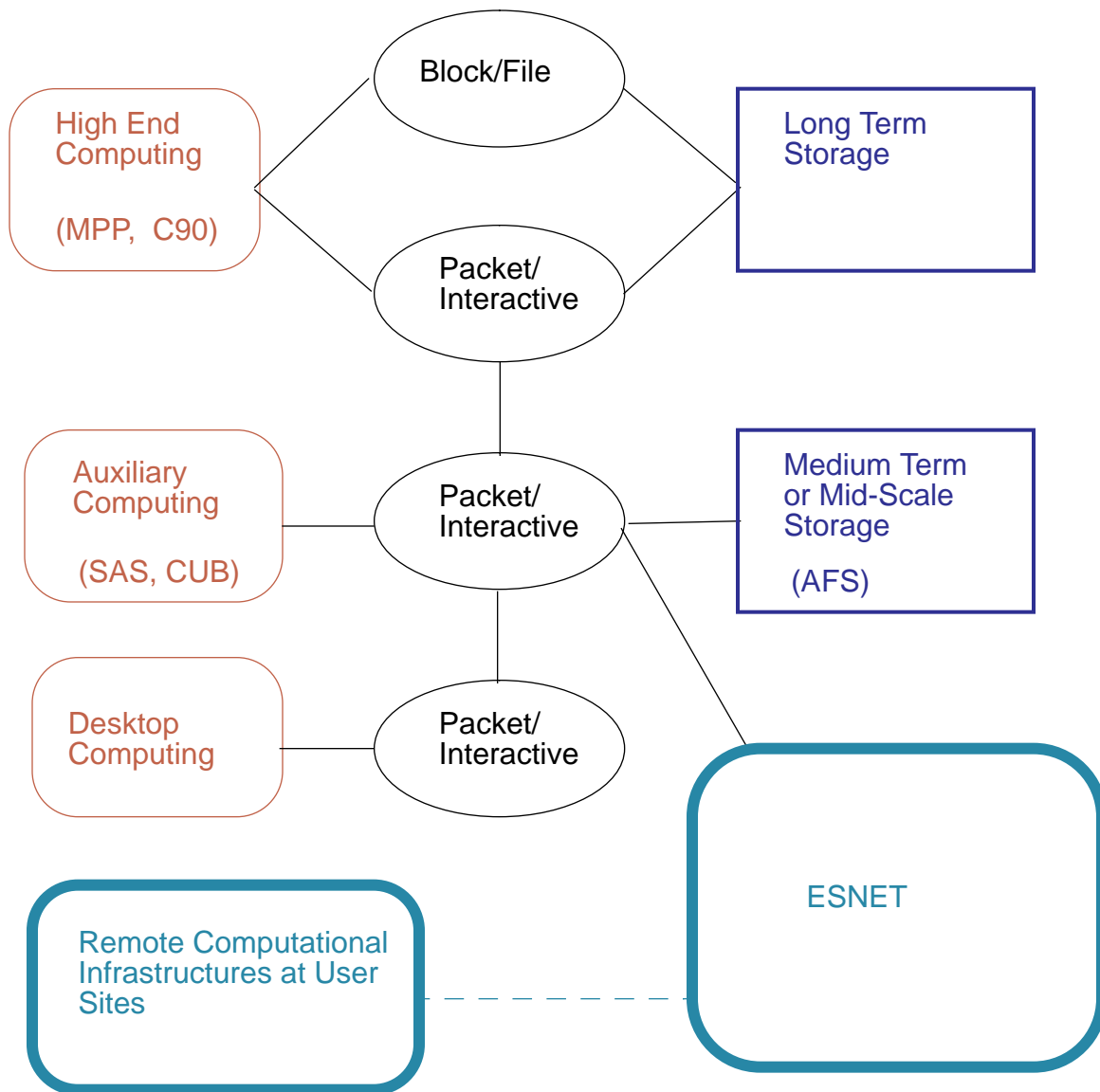
appropriate platform.

4. **NERSC must integrate the remote user's local environment with the centralized NERSC services. This unification of distributed services forms the other part of the UPE.** For users with a rich local environment, the Center must act as the high-end complement to the local environment. *For users with a very basic environment, the Center must seek to provide all the services necessary to compute effectively at the high-end.*

Synopsis of the NERSC Unified Production Environment

In order to provide motivation for the detailed discussion to be found in the Implementation Plan for the UPE, we will provide here an overview of the environment, addressing both the organization of the hardware and the underlying management software which will glue the services together to create the desired environment.

We begin by considering *Figure 3* below (borrowed from the chapter, the **Local Area Network** of the Implementation Plan).



Schematic of Connectivity between Local and Remote Infrastructures

Figure 3

The UPE will coordinate the services seen by the user across the NERSC (centralized) infrastructure seen in the figure above. In the later stages of its implementation, the UPE will extend beyond the centralized environment shown here as we seek to unify the user-local environment with the NERSC centralized environment. The various chapters of the Implementation plan are color-coded in the figure: (1) Development, Computing and Assimilation (in red), (2) Storage (in blue), and (3) the LAN (in black). ESNET and remote

sites are in green. Chapters (4) Administration and (5) User Services represent the software and service glue utilized to unify the environment.

It is more difficult to describe the “unified” part of the “production environment” pictorially. The sense, however, is that the user should be able to access all NERSC services through a common interface after completing a single login process. Each service employed will have access to the same distributed file system. This will allow centralized authentication, shared home directories, and X security via the shared home directories. At a later stage of the development of the unified environment, the user will have the tools required to effectively utilize the resources through distributed batch and interactive computing. The Central User Bank (CUB) can be used for accounting over all services, if necessary. In late phases of the implementation, we will seek to embrace the user’s local environment and unify it with the environment seen at NERSC. The most essential step to achieve the latter is a common file system. However, there are other advantages including centralized licensed software administration, allowing users to check-out temporary licenses for applications codes coming from Independent Software Vendors and run these calculations locally.

We identify in *Table 2* the most important elements of the centralized and distributed management software, together with the target completion dates. For details, see the chapter, **Administration**, of the Implementation Plan.

Service	Type	Completion Date
Integrated administration team/management	Centralized	1995-1H
Single NERSC login/authentication	Centralized	1995-2H
Common banker across all systems	Centralized	1995-1H
Distributed file system including archival storage	Centralized	1995-1996
Distributed batch and interactive computing	Centralized	1995-2H
Licensed Software Administration	Distributed	1995-2H
ER-wide Distributed File System	Distributed	1996-1H

Table 2: Administration Goals for the UPE

An important piece of the UPE puzzle is the configuration and character of the computational hardware. We call this the Development, Computing and Assimilation (DCA) environment, and in *Figure 4* below provide a representation of the **computational hardware** alone. *Each of the three lines represents the cooperating set of computational platforms at a stage in the development of the UPE.* On any given line, any computer standing to the right of the red vertical **stand alone** strip could serve the users without ancillary computational support from other systems. We refer to this as a DCA system.

Any computer to the left of the line represents a system on which the software environment is insufficient to support many users attempting to develop, run and assimilate the results of their work. Therefore, such a computer must be integrated with other systems that can provide the necessary missing services. How well these systems are integrated is a measure of the sophistication of the UPE in this area of service, the DCA environment. Three stages in the evolution of the environment at NERSC are detailed.:

Stage 1 (Lower dashed *Dark Red* line)

By the early 1990's, the Cray-2 environment was augmented by the C90 running Unix and by the SAS system, consisting of HP and SUN workstations. This represented the beginning of a multi-system production environment. **By late 1995**, with the integration of the NERSC Pilot Early Production (PEP) MPP system of about 128-256 processing elements, the UPE will have taken on a new character. Components of the environment (the PEP) are no longer stand-alone from the point of view of DCA environment. Note, the lower line is broken to symbolize the incompleteness of the multisystem UPE at this early stage

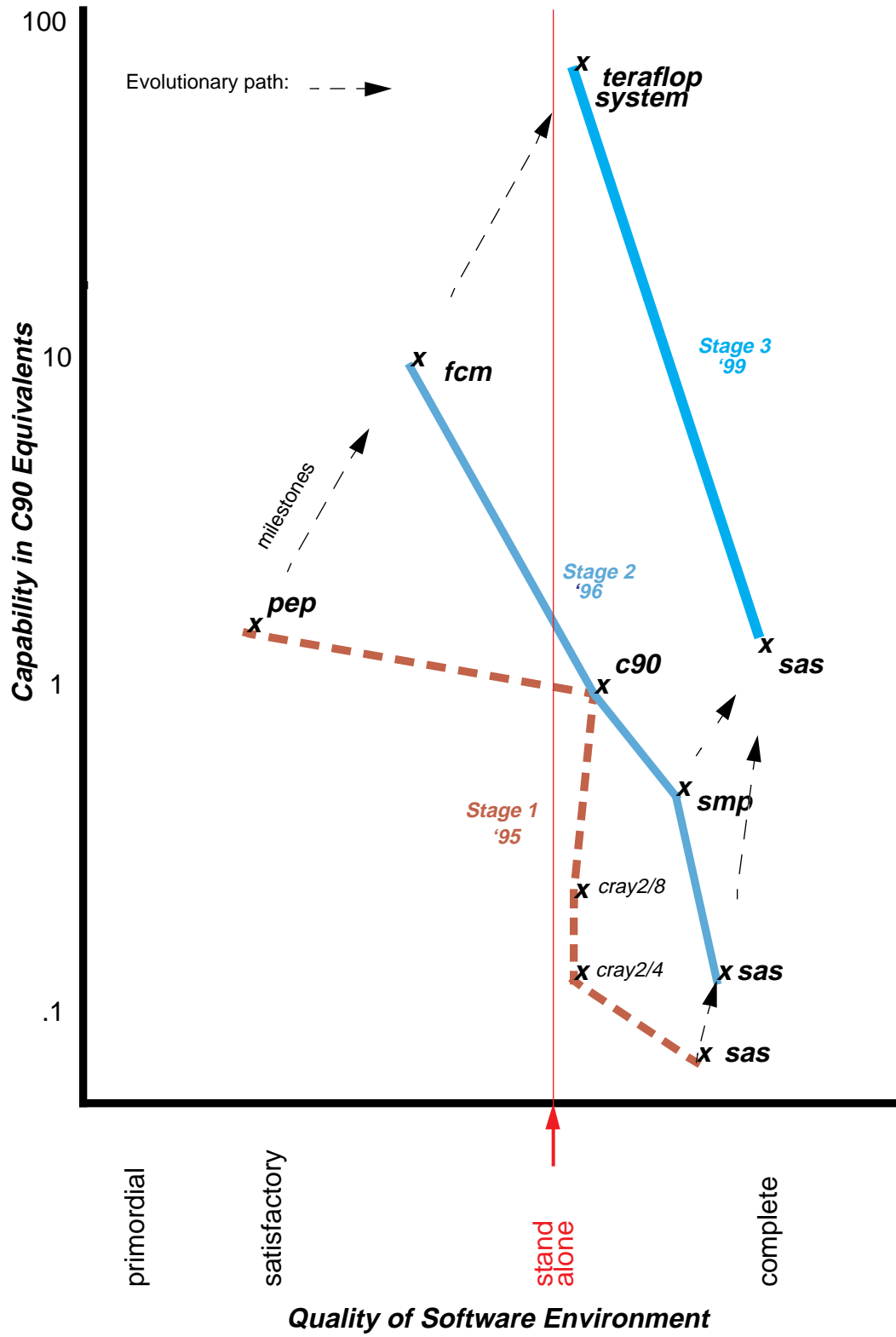
Stage 2 (Middle *Gray* line)

By late 1996, the UPE is significantly improved by two developments, the arrival of the Fully Configured Machine (FCM), an MPP of at least 512 processors, and by the possible addition of at least one SMP. The FCM is acquired only if the "Production Status Requirement" milestones are met on the PEP system. The SMP provides the ability to serve the high end with its rich code development environment and with its post processing abilities. It emulates the FCM on a smaller scale for purposes of debugging applications and limited calibration runs. The FCM has a minimally satisfactory software environment, but cannot serve the user base alone. It benefits from capability partners and requires the SAS (of which the SMP can be considered a part). At this point the environment is integrated across all systems, and can be called a functional UPE. Users will see a single distributed file system.

Stage 3 (Upper *Cyan* line)

By 1998/9, the high-end system will probably not require ancillary computational engines. Nonetheless, it will live in a UPE with a sophisticated, distributed workstation/PC understructure. Thus, the teraflop system will live to the right of the **stand alone** line but will be utilized primarily as a capability engine, other services coming from SAS and from user-local environments tightly coupled into the NERSC environment through the network.

Figure 4: The Three Stages of the DCA Environment



Conclusion

This has provided only the broadest outline of the strategy to be employed by the Center for the next two to three years. The plan focuses on aggressive utilization of new technologies, and the unification of these technologies into a common environment for scientists utilizing this access Center for their computational. research. There is a follow-on document, the Implementation Plan for the Unified Production Environment, which provides many of the details omitted here. Each Chapter focuses on an aspect of the environment, and each chapter begins with motivational material including background information, and follows with the area of primary focus. Then the chapter moves to broad goals (always highlighted in **bold sentences**), and finally concludes with milestones which are specific deliverables and often have associated target completion dates. Both the Strategic Plan and the Implementation Plan will be treated as living documents. They will be modified as necessary as events demand revision, but these documents will continue to represent our current thinking, and they will remind us of where we want to go.